对焦分类方法
The Focusing Classification Method

何沧平
微博人工智能实验室
cangping@staff.weibo.com

2017 年 11 月 16 日

摘要

本文提出一个名为对焦分类的线性分类方法，尝试替代经典的逻辑回归。对焦分类能够从数学论证上保证法向量有界，有直观的几何解释，方便选取更接近最优值的参数初值，在手写数字图像数据集上的分类正确率、收敛速度均显著优于逻辑回归，参数初值即使分类正确率达到了 97.31%。

This paper proposes a new linear classification method named Focusing Classification, with the goal of taking the place of Logistic Regression. Focusing Classification has some advantages: length of its normal vector is limited, intuitional geometrical explanation, parameters' initial values are close to the best values. numerical experiments on the MNIST dataset demonstrate that Focusing Classification has better performance than Logistic Regression on length of its normal vector, accuracy and rate of convergence. With initial parameter values, Focusing Classification gains an accuracy of 97.31%.

关键字: 对焦分类，逻辑回归，线性分类

1 引言

逻辑回归 (Logistic Regression) 是机器学习的一个基础分类方法 [1]。它形式简单，有 LIBLINEAR [2] 这样的现成工具库，工程实现方便，在互联网推荐系统（例如广告点击预测，微博消息推送）中有广泛的应用。但是，逻辑回归仍有一些难以完美解决的的问题。

过拟合现象，即训练一段时间以后，随着训练样本集上的正确率逐渐提高，测试样本上的正确率却不再提高甚至反而下降。过拟合的根本原因尚无共识，目前的应对办法是在损失函数中添加正则化项 [3]，阻止参数变得过大，至于多大算是“过大”，没有具体定义。

虽然正则化缓解了过拟合现象，但它带来了新的麻烦：正则化系数的选择缺少理论指导，只能针对具体训练样本多次试探；正则化还增加了模型复杂度，求解最优化问题需要大量的技巧 [4–11]。

参数初值难选准。逻辑回归的参数只有大致的含义，例如 $\mathbf{w}$ 代表各个特性的权重，b 代表截距。这些含义难以指导选到最优值附近的初值，通常的做法是随机选取初值、预训练。对某个具体训练样本集，按照均匀分布、正态分布等常见分布取值，尝试几次，选用表现最好的分布；先在小样本集上训练，然后将得到的最优参数值用作大样本集上的初值。这种两种方法都费时费力。

本文设计一种名为对焦分类 (Focusing Classification) 的线性二分类方法，克服逻辑回归面临的困难，同时不显著降低分类正确率，尝试代替逻辑回归。在"不让参数过大能够缓解过拟合现象"的假设前提下，对焦分类从数学理论上保证分隔平面法向量的模长有界，从而不必再使用正则化手段来缓解过拟合现象；对焦分类有明确直观的几何意义，方便为参数选取较准确的初值。

本文后续内容这样组织。第 2 节明确二分类问题并提出线性可分这个概念，第 3 节给出逻辑回归的几何解释，第 4 节给出对焦分类的具体公式，第 5 节是对焦分类的几何解释，第 6 节从数学上证明法向量有界，第 7 节给出具体算法实现，第 8 节是数值实验，第 9 节总结全文，第 10 节的附录证明逻辑回归法向量的无限和有界。

2 二分类问题

给定数据集 $D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_m, y_m)\}$，其中 d 为正整数，列向量 $\mathbf{x}_i = (x_{i1}; x_{i2}; \ldots; x_{id})$，$y_i \in \{0, 1\}$。当 $y_i = 1$ 时称 $\mathbf{x}_i$ 是正样本，当 $y_i = 0$ 时称 $\mathbf{x}_i$ 是负样本。二分类问题是要从数据集 D 中学习到一个模型，然后用这个模型预测任意的样本 $\mathbf{x}_j$ 是正样本还是负样本。

对给定数据集 D，记列向量 $\mathbf{w} = (w_1; w_2; \ldots; w_d)$，$\mathbf{c} = (c_1; c_2; \ldots; c_d)$。如果存在一个 d 维平面

$$\mathbf{w}^T(\mathbf{x} - \mathbf{c}) = 0, \quad |\mathbf{w}| \neq 0, \tag{1}$$

使得对任意样本 $\mathbf{x}_i \in D$ 有

$$\begin{cases} y_i = 0, & \text{如果 } \mathbf{w}^T(\mathbf{x}_i - \mathbf{c}) < 0, \\ y_i = 1, & \text{如果 } \mathbf{w}^T(\mathbf{x}_i - \mathbf{c}) \geq 0, \end{cases} \tag{2}$$

或者

$$\begin{cases} y_i = 1, & \text{如果 } \mathbf{w}^T(\mathbf{x}_i - \mathbf{c}) < 0, \\ y_i = 0, & \text{如果 } \mathbf{w}^T(\mathbf{x}_i - \mathbf{c}) \geq 0, \end{cases} \tag{3}$$

那么称数据集 D 是线性可分的。下文仅讨论式 (2) 的情形，式 (3) 情形对应的算法和结论都相同。

如果用平面 (1) 来推测任意样本 $\mathbf{x}_j$ 归属的类别

$$y_j = \begin{cases} 0, & \text{如果 } \mathbf{w}^T(\mathbf{x}_j - \mathbf{c}) < 0, \\ 1, & \text{如果 } \mathbf{w}^T(\mathbf{x}_j - \mathbf{c}) \geq 0. \end{cases}$$

那么 (1) 称为**分隔平面**。如果 $y_j = 0$，那么推测 $\mathbf{x}$ 是负样本；如果 $y_j = 1$，那么推测 $\mathbf{x}_j$ 是正样本。

根据线性可分的定义，任意平面都可以用做分隔平面，区别只是推测效果可能不同。式 (1) 是分隔平面的点法式方程，由解析几何知道，它还有一个等价的斜截式方程。

3 逻辑回归的几何解释与初值

教科书中常以概率的角度讲解逻辑回归。为方便引入对焦分类方法，这里给出逻辑回归的直观几何解释和存在的初值问题。

逻辑回归的目标是从样本集中学习到分隔平面的斜截式方程

$$\mathbf{w}^T\mathbf{x} + b = 0,$$

确定其中的法向量 $\mathbf{w}$ 和截距 b 值。为此用到 Sigmoid 函数

$$\sigma(z) = \frac{1}{1 + e^{-z}},$$

$\sigma(z)$ 的图像如图 1中的红色曲线所示。为了让正样本和负样本分别逼近函数 $\sigma(z)$ 的正负无穷两端，对 $\forall \mathbf{x}_i \in D$，令 $z_i = \mathbf{w}^T\mathbf{x}_i + b$，定义单个样本 $\mathbf{x}_i$ 上的损失函数

$$l(z_i) = \begin{cases} -\ln(1 - \sigma(z_i)), & \text{如果 } y_i = 0, \\ -\ln(\sigma(z_i)), & \text{如果 } y_i = 1. \end{cases}$$

因此，样本集 D 上的损失函数为

$$L(\mathbf{w}, b) = \frac{1}{m}\sum_{i=1}^{m} l(z_i),$$

求解它的最小值

$$\{\hat{\mathbf{w}}, \hat{b}\} = \arg\min_{\mathbf{w},b} \frac{1}{m}\sum_{i=1}^{m} l(z_i). \tag{4}$$

就得到了最优参数 $\hat{\mathbf{w}}$ 和 $\hat{b}$。

损失函数 $l(z_i)$ 能够衡量近似值 $\sigma(z_i)$ 与真实标签 y_i 之间的差距。如图 1 所示，红色曲线是 Sigmoid 函数 $\sigma(z_i)$；在 $\mathbf{x}_i$ 为正样本即 $y_i = 1$ 时，用右侧双向箭头标记的距离反映 $\sigma(z_i)$ 与 $y_i = 1$ 之间距离，z_i 越大，$\sigma(z_i)$ 越接近于 y_i，$l(z_i)$ 越接近于 0（见图 2中红色曲

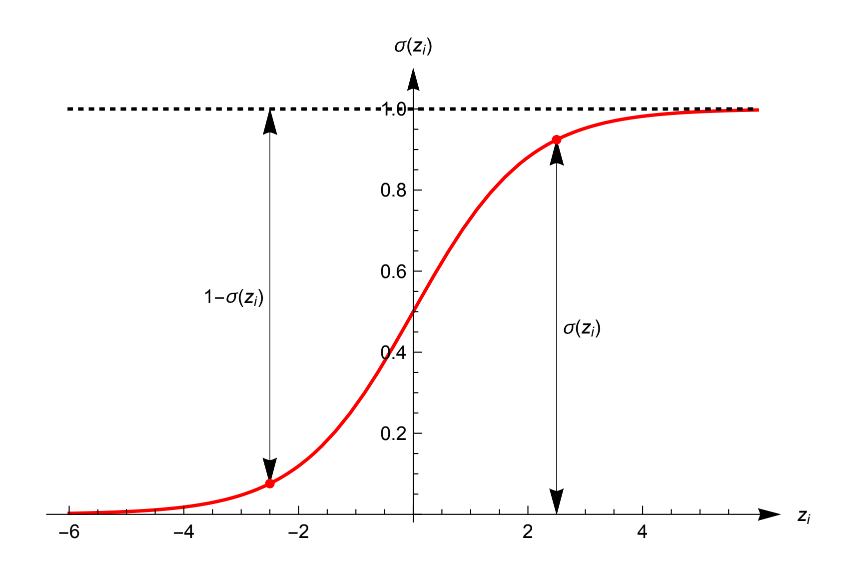


图 1: 逻辑回归：近似值 $\sigma(z_i)$(红色曲线) 与样本标签的距离 (带双向箭头的直线)。

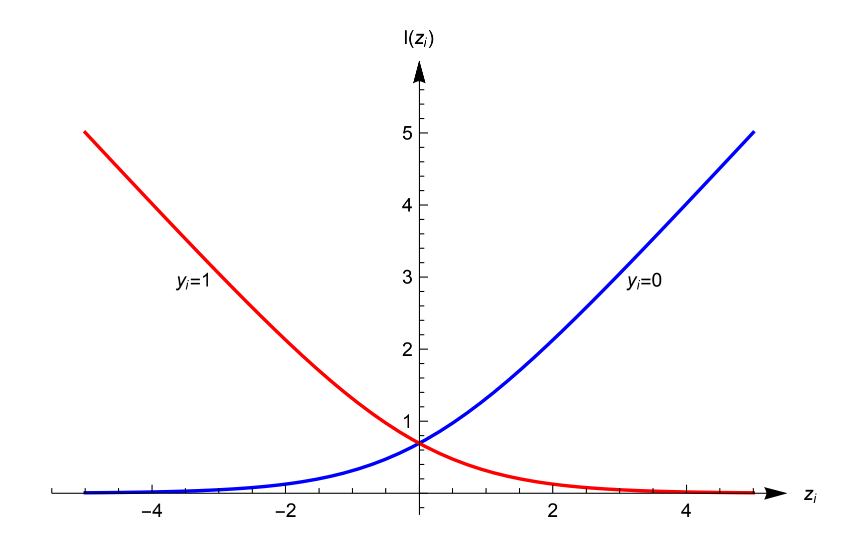


图 2: 逻辑回归：单个样本上的损失函数 $l(z_i)$。

线)；在 $\mathbf{x}_i$ 为负样本即 $y_i=0$ 时，用左侧双向箭头标记的距离反映 $\sigma(z_i)$ 与 $y_i=0$ 之间距离，z_i 越小，$1-\sigma(z_i)$ 越接近于 $1-y_i=1$，$l(z_i)$ 越接近于 0 （见图 2中蓝色曲线)。

图 2画出了单个正样本 (红色) 和单个负样本（蓝色）的损失曲线。直观地理解，如果样本集是线性可分的，那么正样本对应的 z_i 越大，该样本上的损失函数值越小；负样本对应的 z_i 越小，该样本上的损失函数值越小。从而，式 (4) 的计算结果是负样本向 z_i 负无穷方向移动，正样本向 z_i 正无穷方向移动，达到了分类的目的。

虽然能分离正负样本，但逻辑回归还有个问题待解决：初值难选准。

由附录中的定理 4知道，样本集 D 线性可分时，逻辑回归最优分隔平面的法向量 $\hat{\mathbf{w}}$ 的模长是 $+\infty$。这意味着，求解式 (4) 的过程中法向量模长趋向 $+\infty$，但永远无法达到 $+\infty$，必须在适当的时候结束迭代计算。那么问题就来了，不知道结束迭代时法向量的模长是多少，无法为法向量 $\mathbf{w}$ 选择接近最优法向量的初值，导致计算量较多。截距 b 的最优值依赖于法向量 $\mathbf{w}$，因此也难给它一个较好的初值。

由附录中的定理 5知道，样本集 D 线性不可分时，逻辑回归最优分隔平面的法向量 $\hat{\mathbf{w}}$ 有界。虽然不再趋向于 $+\infty$，但仍然无法预先估算 $\hat{\mathbf{w}}$ 的位置。在实际应用中，还会在式 (4) 的目标函数中添加正则化项，导致 $\hat{\mathbf{w}}$ 的取值更加复杂，初值更难选准。

4 对焦分类方法

既然法向量无限是逻辑回归初值难选的一个原因，那么就让对焦分类自动保证法向量有界。逻辑回归采用的斜截式平面方程 (3)，参数含义不直观。对焦分类方法采用斜截式与点法式混合形式的平面方程，使各个参数都有直观的几何意义，指导选取较好的初值。

*对焦变换*的形式为

$$z_i=\mathbf{w}^T(\mathbf{x}_i-\mathbf{c})+b, \tag{5}$$

这里的 $\mathbf{c}$ 是需要指定的任意向量，称为*锚点*；$\mathbf{w}$ 称为*法向量*，实数 b 称为*离心距*，z_i 称为*法向距离*。定义两个实数 F_0 和 F_1，称为焦点，满足 $F_0=-F_1$ 且 $F_1>\ln(3)=1.0986$。记 $G_0=\sigma(F_0), G_1=\sigma(F_1)$。

定义 4 个下标集合 $I_0=\{i|y_i=0,z_i<F_0,i=1,2,\dots,m\}$，$I_1=\{i|y_i=0,z_i\geq F_0,i=1,2,\dots,m\}$，$J_0=\{i|y_i=1,z_i\leq F_1,i=1,2,\dots,m\}$，$J_0=\{i|y_i=1,z_i>F_1,i=1,2,\dots,m\}$。

单个样本 $\mathbf{x}_i\in D$ 上的损失函数定义为

$$h(z_i)=\begin{cases}1-\cos(r(G_0-\sigma(z_i)), & \text{如果 } i\in I_0,\\ 1-\cos(\sigma(z_i)-G_0), & \text{如果 } i\in I_1,\\ 1-\cos(G_1-\sigma(z_i)), & \text{如果 } i\in J_0,\\ 1-\cos(r(\sigma(z_i)-G_1)), & \text{如果 } i\in J_1.\end{cases} \tag{6}$$

这里 r 的取值范围为 $[0, \pi/(2G_1-1)]$，一个典型值是 $r=(1-G_0)/G_0$。将样本集 D 的上损失函数定义为

$$H(\mathbf{w}, b) = \frac{1}{m}\sum_{i=1}^{m} h(z_i). \tag{7}$$

给定锚点 $\mathbf{c}$，在样本集 D 上求损失函数式 (7) 的最小值

$$\{\hat{\mathbf{w}}, \hat{b}\} = \arg\min_{\mathbf{w}, b} H(\mathbf{w}, b) \tag{8}$$

即可得到最优参数 $\hat{\mathbf{w}}$ 和 $\hat{b}$。然后，对任意的样本 $\mathbf{x}_j$，对它做对焦变换

$$z_j = \hat{\mathbf{w}}^T(\mathbf{x}_j - \mathbf{c}) + \hat{b}, \tag{9}$$

用式 (10) 来推测它是正样本或是负样本：

$$y_j = \begin{cases} 0, & \text{如果 } z_j < 0, \\ 1, & \text{如果 } z_j \geq 0, \end{cases} \tag{10}$$

如果 $y_j = 0$，那么推测 $\mathbf{x}_j$ 是负样本；如果 $y_j = 1$，那么推测 $\mathbf{x}_j$ 是正样本。

使用最速下降法等迭代方法求解式 (8) 时，需要计算导数。对 $\forall 1 \leq i \leq m$ 和 $\forall 1 \leq j \leq d$，由式 (5)(6) 知，损失函数的偏导数为

$$h'(z_i) = \begin{cases} -\sin[r(G_0 - \sigma(z_i))]r\sigma'(z_i), & \text{如果 } y_i = 0 \text{且 } z_i < F_0, \\ \sin(\sigma(z_i) - G_0)\sigma'(z_i), & \text{如果 } y_i = 0 \text{且 } z_i \geq F_0, \\ -\sin(G_1 - \sigma(z_i))\sigma'(z_i), & \text{如果 } y_i = 1 \text{且 } z_i \leq F_1, \\ \sin[r(\sigma(z_i) - G_1)]r\sigma'(z_i), & \text{如果 } y_i = 1 \text{且 } z_i > F_1, \end{cases}$$

$$\begin{aligned} \frac{\partial h(z_i)}{\partial w_j} &= \frac{\partial h(z_i)}{\partial z_i}\frac{\partial z_i}{\partial w_j} = \frac{\partial h(z_i)}{\partial z_i}(x_{ij} - c_j), \\ \frac{\partial h(z_i)}{\partial b} &= \frac{\partial h(z_i)}{\partial z_i}\frac{\partial z_i}{\partial b} = \frac{\partial h(z_i)}{\partial z_i}. \end{aligned}$$

从而有

$$\begin{aligned} \frac{\partial H(\mathbf{w}, b)}{\partial \mathbf{w}} = &-\sum_{i\in I_0}\sin[r(G_0 - \sigma(z_i))]r\sigma'(z_i)(\mathbf{x}_i - \mathbf{c}) + \sum_{i\in I_1}\sin(\sigma(z_i) - G_0)\sigma'(z_i)(\mathbf{x}_i - \mathbf{c}) \\ &-\sum_{i\in J_0}\sin(G_1 - \sigma(z_i))\sigma'(z_i)(\mathbf{x}_i - \mathbf{c}) + \sum_{i\in J_1}\sin[r(\sigma(z_i) - G_1)]r\sigma'(z_i)(\mathbf{x}_i - \mathbf{c}), \end{aligned} \tag{11}$$

$$\begin{aligned} \frac{\partial H(\mathbf{w}, b)}{\partial b} = &-\sum_{i\in I_0}\sin[r(G_0 - \sigma(z_i))]r\sigma'(z_i) + \sum_{i\in I_1}\sin(\sigma(z_i) - G_0)\sigma'(z_i) \\ &-\sum_{i\in J_0}\sin(G_1 - \sigma(z_i))\sigma'(z_i) + \sum_{i\in J_1}\sin[r(\sigma(z_i) - G_1)]r\sigma'(z_i). \end{aligned} \tag{12}$$

5 对焦分类的几何解释

本节给出对焦换式 (5) 中的参数 $\mathbf{w}$、$\mathbf{c}$、b 和 z_i 的直观含义，为此引入一个解析几何中的概念。

定义 1. 对 d 维平面 $\mathbf{n}^T(\mathbf{x}-\mathbf{c})=0$ 和 d 维空间的任意点 $\mathbf{x}_i$，称 $\frac{\mathbf{n}^T(\mathbf{x}_i-\mathbf{c})}{|\mathbf{n}|}$ 为点到平面的**有向距离**，称 $\mathbf{n}^T(\mathbf{x}_i-\mathbf{c})$ 为点到平面的**法向距离**。

由定义 1知道，法向距离等于有向距离乘以平面法向量的模长。令 $p_i=\frac{\mathbf{n}^T(\mathbf{x}_i-\mathbf{c})}{|\mathbf{n}|}$，由解析几何知道，如果 $p_i>0$，那么点 $\mathbf{x}_i$ 在法向 $\mathbf{n}$ 方向的一侧（正面）；如果 $p_i<0$，那么点 $\mathbf{x}_i$ 在法向 $\mathbf{n}$ 方向相反的一侧（背面）；如果 $p_i=0$，那么点 $\mathbf{x}_i$ 在平面上。

使用式 (8) 得到的最优参数 $\hat{\mathbf{w}}$ 和 $\hat{b}$ 以后，就得了最优分隔平面的斜截式方程

$$\hat{\mathbf{w}}^T(\mathbf{x}-\mathbf{c})+\hat{b}=0, \tag{13}$$

其中锚点 $\mathbf{c}$ 是任意指定的常向量。假设这个平面的点法式方程为

$$\hat{\mathbf{w}}^T(\mathbf{x}-\hat{\mathbf{c}})=0, \tag{14}$$

那么易知

$$\hat{b}=\hat{\mathbf{w}}^T(\mathbf{c}-\hat{\mathbf{c}}). \tag{15}$$

从而 $\hat{b}$ 的几何意义就是*锚点 $\mathbf{c}$ 到分隔平面的法向距离*，这正是它的名字*离心距*的由来。

由定义 1知道，$z_i=\hat{\mathbf{w}}^T(\mathbf{x}_i-\hat{\mathbf{c}})$ 是点 $\mathbf{x}_i$ 到分隔平面 (14) 的法向距离。对分隔平面来说，只要法向量 $\hat{\mathbf{w}}$ 保持方向不变，模长做任意变化仍然表示同一个平面。注意，零向量没有方向，法向量的模长不能为 0。

实际计算时，能从计算结果中得到分隔平面的斜截式方程 (13)。现在来确定分隔平面的点法式方程 (14) 中的 $\hat{\mathbf{c}}$。注意，在分隔平面固定的情况下，该平面的上任意点均可用作式 (14) 中的 $\hat{\mathbf{c}}$。为了保证唯一性和推导方便，不妨假设 $\mathbf{c}-\hat{\mathbf{c}}$ 与法向量 $\hat{\mathbf{w}}$ 共线，即存在非零实数 s 使得

$$\mathbf{c}-\hat{\mathbf{c}}=s\hat{\mathbf{w}},$$

那么由式 (15) 知

$$\hat{b}=\hat{\mathbf{w}}^T s\hat{\mathbf{w}}=s|\hat{\mathbf{w}}|^2,$$

从而有

$$|\mathbf{c}-\hat{\mathbf{c}}|=s|\hat{\mathbf{w}}|=\frac{\hat{b}}{|\hat{\mathbf{w}}|}, \tag{16}$$

$$\hat{\mathbf{c}}=\mathbf{c}-s\hat{\mathbf{w}}=\mathbf{c}-\frac{\hat{b}}{|\hat{\mathbf{w}}|^2}\hat{\mathbf{w}}.$$

以 2 维样本为例说明各个参数的含义。对 2 维样本，分隔平面退化为直线。图 3中，黑色直线 $\mathbf{w}^T(\mathbf{x}-\mathbf{c})=0$ 是迭代计算的起点，法向量 $\mathbf{w}$ 和锚点 $\mathbf{c}$ 随意取值，离心距 b 的初值

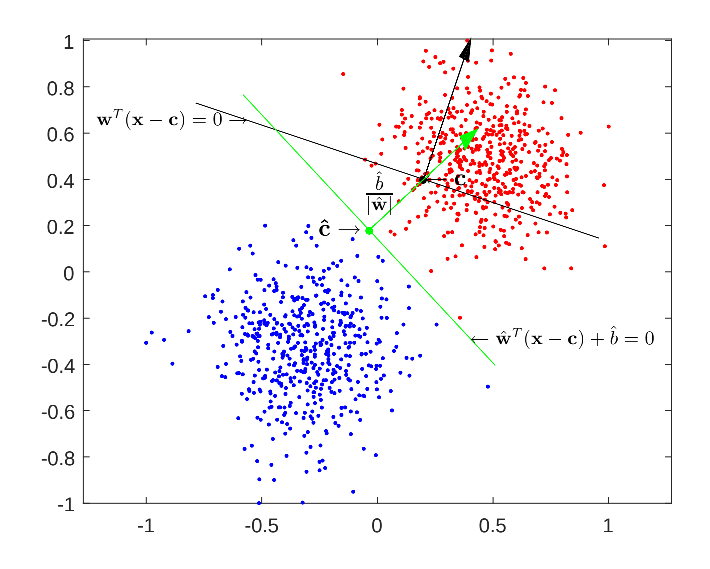


图 3: 对焦分类中参数的几何意义

取 0。带箭头的黑色直线是法向量 $\mathbf{w}$，两条直线的交叉点是 $\mathbf{c}$。绿色直线是迭代若干次之后得到的最优分隔线 $\hat{\mathbf{w}}^T(\mathbf{x}-\mathbf{c})+\hat{b}=0$，带箭头的绿色直线是它的法向 $\hat{\mathbf{w}}$，两条直线的交叉点就是 $\hat{\mathbf{c}}$。点 $\mathbf{c}$ 和点 $\hat{\mathbf{c}}$ 之间距离为 $\hat{b}/|\hat{\mathbf{w}}|$，对应式 (16)，这意味着 $\hat{b}$ 能够衡量人为指定的锚点 $\mathbf{c}$ 与最优锚点 $\hat{\mathbf{c}}$ 之间的距离。

现在考察对焦对类方法的分类机理。

观察损失函数 $h(z_i)$ 的走势，见图 4，$y_i=1$ 标识的曲线对应正样本，$y_i=0$ 标识的曲线对应负样本。由图 4和式 (6) 可以看出，在 $r=(1-G_0)/G_0$ 时，正负样本的损失曲线关于纵轴轴对称，两条曲线在焦点 F_1 和 F_0 处的值分别为 0，且有

$$\lim_{\substack{y_i=0\\ z_i\to-\infty}} h(z_i)=\lim_{\substack{y_i=0\\ z_i\to+\infty}} h(z_i)=\lim_{\substack{y_i=1\\ z_i\to-\infty}} h(z_i)=\lim_{\substack{y_i=1\\ z_i\to+\infty}} h(z_i)=1-\cos(G_1).$$

从图 4可以看出，也可以用式 (6) 严格证明：在 $y_i=0$ 时，$h(z_i)$ 在 $(-\infty,F_0]$ 严格单调递减，在 $[F_0,+\infty)$ 严格单调递增；在 $y_i=1$ 时，$h(z_i)$ 在 $(-\infty,F_1]$ 严格单调递减，在 $[F_1,+\infty)$ 严格单调递增。

对一个正样本 $\mathbf{x}_i$ 来说，它的法向距离 z_i 落在焦点 F_1 上时对整体损失 $H(\mathbf{w},b)$ 的贡献最小。但是，所有正样本的法向距离 z_i 通常会散落在一个区间中，不会全部落在焦点 F_1 上。因此，调整对焦变换中的参数使所在正样本的法向距离聚集在焦点 F_1 附近，就能促使整体损失 $H(\mathbf{w},b)$ 达到最小值。同理，负样本的法向距离集中在焦点 F_0 附近时，也能促进整体损失 $H(\mathbf{w},b)$ 达到最小值。

观察损失函数导数 $h'(z)$ 的图像，见图 5，两条曲均连续，焦点 F_0、F_1 之外均光滑。这

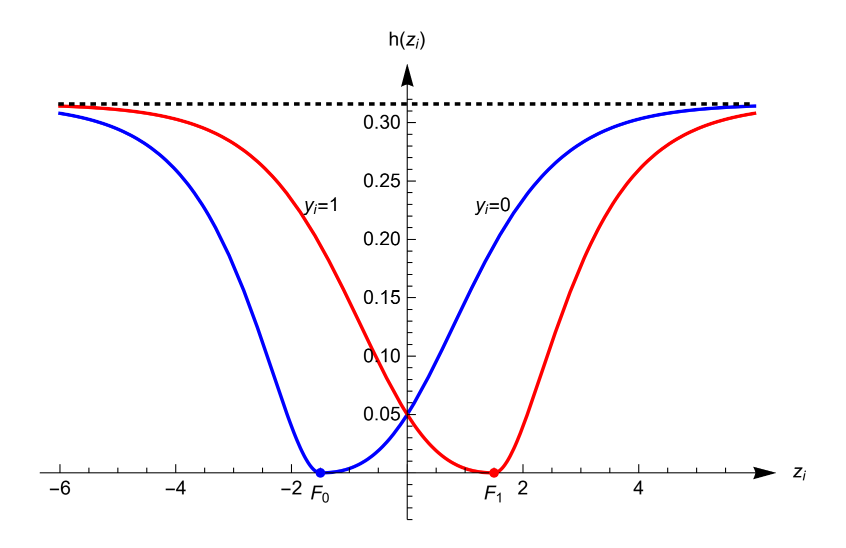


图 4: 单个样本上的损失函数 $h(z_i)$，焦点 $F_0 = -1.5$，$F_1 = 1.5$, $r = (1 - G_0)/G_0$。

意味着迭代过程会平稳地收敛到最优点，不会出现前进方向的跳变。越靠近焦点，导数的绝对值越大，这意味着迭代计算时将加速逼近最优值，不会在最优值附近震荡。沿着正无穷和负无穷方向，导数快速逼近 0，这意味着如果对焦变换参数的初值偏离最优值太远，那么收敛会很慢。不过，使用时不必担心这个问题，7.1节给出的方法能选到接近最优值的初值。

6 法向量边界

本节证明对焦变换中的最优法向量 $\hat{\mathbf{w}}$ 和 $\hat{b}$ 是有界的，从而不必担心逻辑回归中出现的法向无限问题。先证明 1 维样本的情形，再证明 d 维样本的情形。

6.1 1 维样本

对 1 维样本，样本向量在形式上就成为 $\mathbf{x}_i = (x_{i1})$，简记为 x_i，将 $\mathbf{w} = (w_1)$ 简记为 w，将 $\mathbf{c} = (c_1)$ 简记为 c。假设样本集 $D_1 = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$ 是线性可分的，即存在一个点 $\hat{c}$ 使得

$$\begin{cases} y_i = 0, & \text{如果 } x_i < \hat{c}, \\ y_i = 1, & \text{如果 } x_i > \hat{c}, \end{cases} \tag{17}$$

或者

$$\begin{cases} y_i = 1, & \text{如果 } x_i < \hat{c}, \\ y_i = 0, & \text{如果 } x_i > \hat{c}. \end{cases} \tag{18}$$

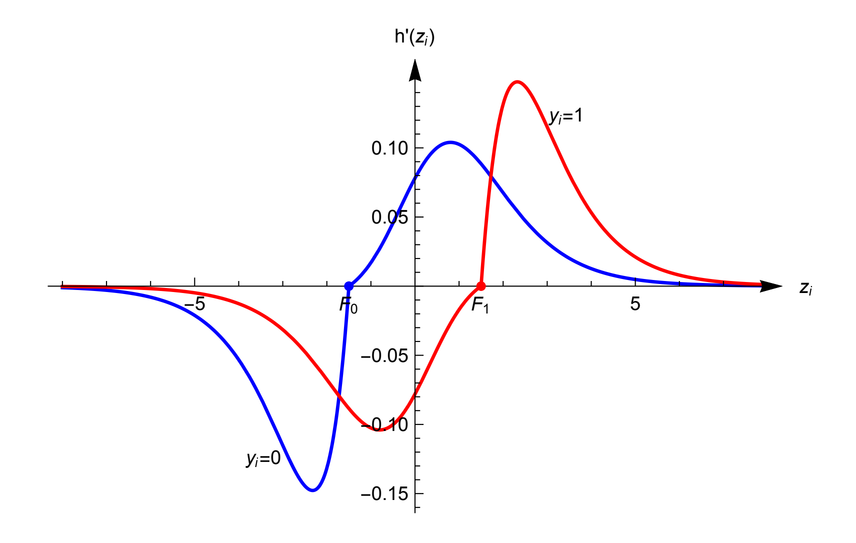


图 5: 单个样本上的损失函数的导数 $h'(z_i)$，焦点 $F_0 = -1.5$，$F_1 = 1.5$，$r = (1 - G_0)/G_0$。

这时最优分隔平面退化一个点：$z = \hat{c}$。不失一般性，再假设 $\hat{c} = 0$。如果 $\hat{c} \neq 0$，只需对所有样本 $x_i, i = 1, 2, \ldots, m$ 都作一个平移操作，不影响分类结果。$\hat{c} = 0$ 时，取 $c = \hat{c}$，由式 (5) 知 $\hat{b} = 0$，对焦变换就成为

$$z_i = w(x_i - \hat{c}) + \hat{b} = wx_i. \tag{19}$$

既然 $\hat{b}$ 已经确定，现在寻找最优参数的工作只剩下确定式 (19) 中的最优标量 $\hat{w}$。

接下来仅讨论 (17) 成立的情形，(18) 成立的情形证明方法相同。

定理 1. *1 维样本集 D 线性可分时，对焦分类最优分隔平面的法向量有界。*

证: 为方便叙述，假设

$$x_{m_1+m_0} \leq x_{m_1+m_0-1} \leq \cdots \leq x_{m_1+1} < 0 < x_1 \leq x_2 \leq \cdots \leq x_{m_1},$$

其中 m_0 和 m_1 为正整数，且满足 $m_0 + m_1 = m$。记

$$h_{00}(z_i) = \begin{cases} 1 - \cos(r(G_0 - \sigma(z_i))), & \text{如果 } y_i = 0 \text{且 } z_i < F_0, \\ 0, & \text{其它情形}, \end{cases}$$

$$h_{01}(z_i) = \begin{cases} 1 - \cos(\sigma(z_i) - G_0), & \text{如果 } y_i = 0 \text{且 } z_i \geq F_0, \\ 0, & \text{其它情形}, \end{cases}$$

$$h_{10}(z_i) = \begin{cases} 1 - \cos(G_1 - \sigma(z_i)), & \text{如果 } y_i = 1 \text{且 } z_i \leq F_1, \\ 0, & \text{其它情形}, \end{cases} \tag{20}$$

$$h_{11}(z_i) = \begin{cases} 1 - \cos(r(\sigma(z_i) - G_1)), & \text{如果 } y_i = 1 \text{且 } z_i > F_1, \\ 0, & \text{其它情形，} \end{cases} \tag{21}$$

$$H_0(w) = \sum_{y_i=0} h(z_i), \ H_1(w) = \sum_{y_i=1}^{m} h(z_i), \tag{22}$$

从而式 (7) 变为

$$H(w,b) = \frac{1}{m}\left[H_0(w) + H_1(w)\right]. \tag{23}$$

在 $wx_{m_1} < F_1$ 即 $w < F_1/x_{m_1}$ 时，由式 (19)(20)(22) 得到

$$H_1(w) = \sum_{i=1}^{m_1} h_{10}(wx_i), \quad H_1'(w) = x_i \sum_{i=1}^{m_1} h_{10}'(wx_i),$$

对 $1 \le i \le m_1$，$x_i > 0$，$h_{10}'(wx_i) < 0$，因此 $H_1'(w) < 0$，$H_1(w)$ 在 $(-\infty, F_1/x_{m_1}]$ 严格单调递减。

在 $wx_1 > F_1$ 即 $w > F_1/x_1$ 时，由式 (19)(21)(22) 得到

$$H_1(w) = \sum_{i=1}^{m_1} h_{11}(wx_i), \quad H_1'(w) = x_i \sum_{i=1}^{m_1} h_{11}'(wx_i),$$

对 $1 \le i \le m_1$，$x_i > 0$，$h_{11}'(wx_i) > 0$，因此 $H_1'(w) > 0$，$H_1(w)$ 在 $[F_1/x_1, \infty)$ 严格单调递增。

综合 $H_1(w)$ 在 $w < F_1/x_{m_1}$ 和 $w > F_1/x_1$ 两种情形下的单调性，再考虑到 $H_1(w)$ 一阶连续可导，可知 $H_1(w)$ 的全局最小点 $\hat{w}_1$ 落在区间 $[F_1/x_{m_1}, F_1/x_1]$。

用同样的方法可以证明，$H_0(w)$ 的全局最小点 $\hat{w}_0$ 落在区间 $[F_0/x_{m_1+m_0}, F_0/x_{m_1+1}]$。由式 (23) 知，$H(w,b)$ 的全局最小点 $\hat{w}$ 落在区间 $[\min\{F_0/x_{m_1+m_0}, F_1/x_{m_1}\}, \max\{F_1/x_1, F_0/x_{m_1+1}\}]$，即 $\hat{w}$ 有界。

[证毕]

6.2 d 维样本

借助法向距离可以将 $d \ge 2$ 维样本集 $D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_m, y_m)\}$ 上的二分类问题转化为 1 维样本集上的二分类问题，从而完成证明。

定理 2. *d 维样本集线性可分时，对焦分类最优分隔平面的法向量有界。*

证： 假设 d 维样本集 D 是线性可分的，再假设将 D 完全正确分类的最优分隔平面为

$$\hat{\mathbf{w}}^T(\mathbf{x} - \mathbf{c}) + \hat{b} = 0. \tag{24}$$

令单位向量 $\mathbf{n} = \hat{\mathbf{w}}/|\hat{\mathbf{w}}|$，$p_i = \mathbf{n}^T(\mathbf{x}_i - \mathbf{c}) + \hat{b}$，显然 p_i 是样本 $\mathbf{x}_i$ 到平面 (24) 的有向距离，样本集 D 上二分类问题等价于 1 维样本集 $D_1 = \{(p_1, y_1), (p_2, y_2), \ldots, (p_m, y_m)\}$ 上的二分类问题，假设 D_1 上的对焦变换为

$$z_i = \lambda p_i. \tag{25}$$

由定理 1知，式 (25) 的最优参数 $\hat{\lambda}$ 有界，从而样本集 D 上的最优法向量 $\hat{\mathbf{w}} = \hat{\lambda}\mathbf{n}$ 有界。

[证毕]

6.3 样本集线性不可分

通常情况下，样本集是线性不可分的，即对任意给定的分隔平面，其正面（或背面）都有一些负样本 (或正样本)。这种情形下，最优分隔平面法向量有界的前提条件变得复杂。

观察图 4中的损失函数曲线，猜测会有一个暂无证明的模糊结论：

定理 3. *d 维样本集 D 线性不可分时，如果所有正样本的中心点与所有负样本的中心点显著分离，那么对焦分类的最优法向量有界。*

7 对焦分类的算法实现

本节起，考虑一般性的 $d \geq 1$ 维样本集，为叙述便利，统一采用向量符号，不再单独考虑 $d = 1$ 的情形。

7.1 参数初值

用最速下降法等方法迭代求解式 (8) 时，需要给出式 (5) 中的 3 个参数 $\mathbf{c}$、b 和 $\mathbf{w}$ 的初值。

由图 6直观地看，正负样本都是聚集在某个区域。由统计规律知道，很多随机事件服从正态分布。对正态分布而言，数学期望是它的中心，正样本中心和负样本中心之间的中点应该位于分隔平面的附近。定义 2 个集合 $K_0 = \{i | y_i = 0, 1 \leq i \leq m\}$ 和 $K_1 = \{i | y_i = 1, 1 \leq i \leq m\}$，将 K_0 和 K_1 中元素的数量分别记为 m_0 和 m_1。将正样中心和负样本中心分别记为

$$\boldsymbol{\mu}_1 = \frac{1}{m_1} \sum_{i \in K_1} \mathbf{x}_i,\ \boldsymbol{\mu}_0 = \frac{1}{m_0} \sum_{i \in K_0} \mathbf{x}_i, \tag{26}$$

从而锚点 $\mathbf{c}$ 可以指定为

$$\mathbf{c} = \frac{\boldsymbol{\mu}_0 + \boldsymbol{\mu}_1}{2}. \tag{27}$$

理想情况下，$\mathbf{c}$ 恰好落在最优分隔平面上，此时有 $\mathbf{c} = \hat{\mathbf{c}}$，由式 (16) 知道 $\hat{b} = 0$。因此，b 的初值应选为

$$b = 0.$$

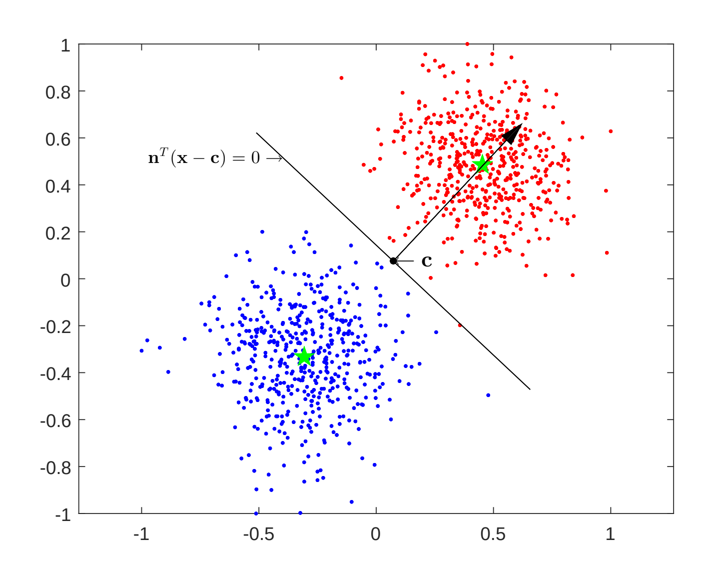


图 6: 参数的初值。2 个五角星分别为正负样本的中心。

由图 4知，正负样本会分别聚集在焦点 F_1、F_0 附近，因此选取 $\mathbf{w}$ 的初值使得正样本的中心 $\boldsymbol{\mu}_1$ 落在焦点 F_1 上，使得负样本的中心 $\boldsymbol{\mu}_0$ 落在焦点 F_0 上。几乎不存 $\mathbf{w}$ 使得正负样本中心同时落在焦点上，因此实际操作中选取一个中间值。这个初值几乎肯定不是最优值，但离最优值不会太远。

为了确定法向量 $\mathbf{w}$ 的初值，先确定其方向 $\mathbf{n}$，再确定其模长 λ，即 $\mathbf{w}=\lambda\mathbf{n}$。用正负样本中心的连线方向作为 $\mathbf{n}$ 的方向，即

$$\mathbf{n}=(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)/|\boldsymbol{\mu}_1-\boldsymbol{\mu}_0|. \tag{28}$$

令 $\theta_0=\mathbf{n}^T(\boldsymbol{\mu}_0-\mathbf{c})$，$\theta_1=\mathbf{n}^T(\boldsymbol{\mu}_1-\mathbf{c})$，

$$\lambda=\frac{1}{2}\left(\frac{F_0}{\theta_0}+\frac{F_1}{\theta_1}\right),$$

从而法向量的初值就为

$$\begin{aligned}\mathbf{w} &= \lambda\mathbf{n}\\ &= \frac{1}{2}\left(\frac{F_0}{\theta_0}+\frac{F_1}{\theta_1}\right)\frac{\boldsymbol{\mu}_1-\boldsymbol{\mu}_0}{|\boldsymbol{\mu}_1-\boldsymbol{\mu}_0|}.\end{aligned} \tag{29}$$

7.2 调整标准差

在实际应用案例中，通常正样本和负样本的方差不相等，甚至相差很大，这种差别导致对焦分类的分类精度稍低于逻辑回归，原因也相当直观。

实际案例中的样本集通常是几乎线性可分的：只有少量的样本（例如小于 5%）被错误分类，因而逻辑回归法向量的模长相当大，正负样本对应的 z_i 均远离原点。从图 1可以看出，在远离原点的地方，例如正无穷方向，不同样本的损失函数值的差别很小，远小于在原点附近的差别，从而逻辑回归对方差差异不敏感。

对焦回归就不一样了，大量样本聚集在焦点附近，样本之间的损失函数值差别大得多，因此必须调整均衡。

利用 (28) 中的单位向量 $\mathbf{n}$，令 $p_i = \mathbf{n}^T\mathbf{x}_i$。定义集合 $P_0 = \{p_i | i \in K_0\}$, $P_1 = \{p_i | i \in K_0\}$，将 P_0 中所有元素的标准差记为 v_0，将 P_1 中所有元素的标准差记为 v_1。定义常量*标准差比例*为

$$\eta = \begin{cases} v_0/v_1, & \text{如果} v_0 \le v_1, \\ v_1/v_0, & \text{如果} v_0 > v_1. \end{cases} \tag{30}$$

记

$$\xi_0 = \mathbf{w}^T(\boldsymbol{\mu}_0 - \mathbf{c}) + b, \quad \xi_1 = \mathbf{w}^T(\boldsymbol{\mu}_1 - \mathbf{c}) + b. \tag{31}$$

现在可以对 z_i 做标准差修正了。当 $v_0 \le v_1$ 时，令

$$\bar{z}_i = \begin{cases} z_i, & \text{如果} i \in K_0, \\ \eta(z_i - \xi_1) + \xi_1, & \text{如果} i \in K_1, \end{cases} \tag{32}$$

当 $v_0 > v_1$ 时，令

$$\bar{z}_i = \begin{cases} \eta(z_i - \xi_0) + \xi_0, & \text{如果} i \in K_0, \\ z_i, & \text{如果} i \in K_1. \end{cases} \tag{33}$$

在后续迭代计算中，用 $\bar{z}_i$ 代替损失函数和损失函数导数公式里的 z_i。

7.3 迭代算法

这里用最速下降法寻找对焦分类的最优参数，即求解最小化问题式 (8)。使用其它最优法算法时，请参照实施。

步 1，获取数据：样本集 $D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_m, y_m)\}$。

步 2，离心距赋初值：$b = 0$。

步 3，锚点赋初值：使用式 (26) 计算正负样本的中心点 $\boldsymbol{\mu}_1$ 和 $\boldsymbol{\mu}_0$，使用式 (27) 得到锚点初值 $\mathbf{c} = (\boldsymbol{\mu}_0 + \boldsymbol{\mu}_1)/2$。

步 4，指定焦点 F_0、F_1。按照第 4节的规定，F_1 的取值范围是 $(\ln(3), +\infty)$，一个典型值是 $F_1 = 1.5$；指定系数 r 的值，按照第 4节的规定，r 的取值范围是 $[0, \pi/(2G_1 - 1)]$，一个典型值是 $r = (1 - G_0)/G_0$。指定最速下降法的下降步长 s，这个值可以随意指定，例如 $s = 0.1$，但不可太大，否则可能导致迭代不收敛。指定收敛阈值 τ，这个值也可以随意指定为一个正数，例如 $\tau = 10^{-5}$。令迭代计数器 $k = 0$，损失值 Loss0=1。

步 5，法向量赋初值：用式 (28) 得到单位向量 $\mathbf{n}$，用式 (29) 给法向量 $\mathbf{w}$ 指定初值。

步 6，用式 (30) 计算标准差比例 η。

现在开始最速度下降法迭代。

步 7，对样本集 D 中的所有样本作对焦变换，即用 (5) 计算 $z_i = \mathbf{w}^T(\mathbf{x}_i - \mathbf{c}) + b, i = 1, 2, \ldots, m$。

步 8，用式 (32) 或 (33) 修正 $\{z_i | i = 1, 2, \ldots, m\}$ 得到 $\{\bar{z}_i | i = 1, 2, \ldots, m\}$。

步 9，计算损失函数值：用 $\bar{z}_i$ 代替式 (7) 中的 z_i，记 $\text{Loss1} = H(\mathbf{w}, b)$。如果 $\text{Loss0} - \text{Loss1} < \tau$，那么跳转至步 11；否则，令 $\text{Loss0} = \text{Loss1}$。

步 10，参数更新，使用式 (11)(12) 计算

$$\mathbf{w} = \mathbf{w} + s\frac{\partial H(\mathbf{w}, b)}{\partial \mathbf{w}}, \quad b = b + s\frac{\partial H(\mathbf{w}, b)}{\partial b},$$

跳转至步 7。

步 11，计算样本集 D 上的分类正确率：对任意样本 $\mathbf{x}_j$，用 (9) 计算 z_j，然后用式 (32) 或 (33) 做标准差计整得到 $\bar{z}_j$，用式 (10) 推测样本 $\mathbf{x}_j$ 是正样本或负样本；统计所有样本的推测结果，得到正确率。结束。

8 数值实验

对焦回归的设计目标是解决逻辑回归所面临的问题，因此本节对比二者的分类正确率、初值优劣、收敛速度、法向量模长。

MNIST [12] 数据库是典型的分类数据集，它包含数字 0~9 的手写图像，6 万个图像用于训练，1 万个图像用于测试。将数字 0~9 的图像分别抽取出来，形成 10 个训练子集和 10 个测试子集。

由 MNIST 官网知道，目前还没有方法能将 MNIST 完全正确地分类，从而可以认为 MNIST 不是线性可分的。由定理 5知道，这种情形下逻辑回归的法向量有界，因此不使用正则化措施。最速下降法的下降步长指定为 0.1，迭代次数设为 2000，法向量 $\mathbf{w}$ 的每个元素都按照服从均匀分布 $U(-1/784, 1/784)$ 来选取初值，截距 b 按照服从均匀分布 $U(-1, 1)$ 来选取初值。

用对焦分类计算时，指定焦点为 $F_1 = 1.5$，$F_0 = -1.5$，$r = (1 - G_0)/G_0$，指定最速下降法的下降步长为 $s = 0.1$。

8.1 分类正确率

用分类正确的样本数量除以样本总量就得到分类*正确率*。逻辑回归在训练集上的正确率见表 1，正确率均值为 98.7191%；逻辑回归在测试集上的正确率见表 2，正确率均值为 98.6730%。

表 1: 逻辑回归在 MNIST 训练集上的正确率%

负-正	0	1	2	3	4	5	6	7	8
1	99.87	-	-	-	-	-	-	-	-
2	99.00	99.08	-	-	-	-	-	-	-
3	99.43	99.09	97.38	-	-	-	-	-	-
4	99.65	99.62	98.58	99.46	-	-	-	-	-
5	98.66	99.48	98.10	96.03	98.99	-	-	-	-
6	99.21	99.76	98.36	99.48	99.12	98.08	-	-	-
7	99.69	99.48	98.69	98.69	98.88	99.33	99.86	-	-
8	99.29	98.21	97.46	97.02	99.26	96.27	98.99	99.06	-
9	99.52	99.52	98.80	98.37	96.89	98.70	99.81	95.91	98.24

表 2: 逻辑回归在 MNIST 测试集上的正确率%

负-正	0	1	2	3	4	5	6	7	8
1	99.95	-	-	-	-	-	-	-	-
2	99.11	99.31	-	-	-	-	-	-	-
3	99.75	99.63	97.65	-	-	-	-	-	-
4	99.85	99.91	98.46	99.60	-	-	-	-	-
5	98.99	99.61	97.97	96.27	99.20	-	-	-	-
6	98.97	99.57	98.34	99.39	98.81	98.05	-	-	-
7	99.60	99.21	97.96	98.04	98.91	99.22	99.60	-	-
8	99.39	98.91	97.31	96.93	99.34	95.82	98.81	98.10	-
9	99.30	99.53	98.68	98.27	97.04	98.42	99.69	96.07	97.78

这些表格里，第 1 行中的数字 0~8 表示正样本对应的数字，第 1 列中的数字 1~9 表示负样本对应的数字。+1 对角线上的位置表示正负样本用同一个数字的图像，无意义，不需要计算；右上角部分与左下角部分对称，故不再列出数值，用"-"代替。

对焦分类在训练集上的正确率见表 3，正确率均值为为 99.0204%；对焦分类在测试集上的正确率见表 4，正确率均值为 99.0825%。

用对焦分类正确率减去逻辑回归的正确率，得到表 5和表 6。在训练集样本上和测试样本集上，对焦分类的平均正确率比逻辑回归的平均正确率分别高 0.3007% 和 0.4095%。

8.2 初值优劣

逻辑回归的初始值随机选取，导致初始正确率不高，见表 7，正确率均值为 49.27%，等效于随机推测样本是正样本或是负样本。

表 3: 对焦分类在 MNIST 训练集上的正确率%

负-正	0	1	2	3	4	5	6	7	8
1	99.98	-	-	-	-	-	-	-	-
2	99.34	99.61	-	-	-	-	-	-	-
3	99.76	99.68	97.12	-	-	-	-	-	-
4	99.96	99.88	99.65	99.72	-	-	-	-	-
5	99.77	99.62	98.42	95.80	99.24	-	-	-	-
6	99.70	99.96	98.51	99.15	99.29	97.99	-	-	-
7	99.98	99.91	99.21	99.31	99.12	99.49	99.89	-	-
8	99.67	99.63	98.25	96.44	99.48	98.86	98.34	98.65	-
9	99.72	99.79	99.48	98.95	96.89	98.29	99.79	96.09	98.53

表 4: 对焦分类在 MNIST 测试集上的正确率%

负-正	0	1	2	3	4	5	6	7	8
1	100.00	-	-	-	-	-	-	-	-
2	99.25	99.77	-	-	-	-	-	-	-
3	99.90	99.86	97.99	-	-	-	-	-	-
4	99.95	100.00	99.70	99.80	-	-	-	-	-
5	99.79	99.85	98.80	96.90	99.41	-	-	-	-
6	99.90	99.90	98.34	99.49	99.28	97.73	-	-	-
7	99.85	100.00	98.64	98.77	99.30	99.43	99.85	-	-
8	99.69	99.86	98.65	96.62	99.39	98.66	98.40	98.15	-
9	99.65	99.95	99.41	98.96	97.09	98.32	99.75	96.37	98.34

表 5: 对焦分类正确率减去逻辑回归正确率%（训练）

负-正	0	1	2	3	4	5	6	7	8
1	0.1184	-	-	-	-	-	-	-	-
2	0.3367	0.5354	-	-	-	-	-	-	-
3	0.3318	0.5904	-0.2564	-	-	-	-	-	-
4	0.3060	0.2622	1.0678	0.2673	-	-	-	-	-
5	1.1107	0.1398	0.3164	-0.2251	0.2486	-	-	-	-
6	0.4898	0.1975	0.1516	-0.3237	0.1616	-0.0882	-	-	-
7	0.2954	0.4305	0.5236	0.6131	0.2478	0.1540	0.0246	-	-
8	0.3822	1.4214	0.7875	-0.5759	0.2138	2.5905	-0.6458	-0.4127	-
9	0.2022	0.2679	0.6803	0.5795	0	-0.4046	-0.0253	0.1801	0.2881

表 6: 对焦分类正确率减去逻辑回归正确率%（测试）

负-正	0	1	2	3	4	5	6	7	8
1	0.0473	-	-	-	-	-	-	-	-
2	0.1491	0.4615	-	-	-	-	-	-	-
3	0.1508	0.2331	0.3428	-	-	-	-	-	-
4	0.1019	0.0945	1.2413	0.2008	-	-	-	-	-
5	0.8013	0.2467	0.8316	0.6309	0.2134	-	-	-	-
6	0.9288	0.3344	0	0.1016	0.4639	-0.3243	-	-	-
7	0.2490	0.7859	0.6796	0.7360	0.3980	0.2083	0.2518	-	-
8	0.3071	0.9483	1.3460	-0.3024	0.0511	2.8403	-0.4141	0.0499	-
9	0.3519	0.4198	0.7349	0.6934	0.0502	-0.1052	0.0508	0.2946	0.5547

表 7: 逻辑回归在训练集上的初始正确率%

负-正	0	1	2	3	4	5	6	7	8
1	46.77	-	-	-	-	-	-	-	-
2	49.85	46.91	-	-	-	-	-	-	-
3	50.86	47.63	49.28	-	-	-	-	-	-
4	49.66	46.42	50.49	48.79	-	-	-	-	-
5	52.21	44.57	47.64	46.93	51.87	-	-	-	-
6	45.95	46.75	50.17	49.12	49.68	47.81	-	-	-
7	51.16	51.83	48.74	49.46	48.25	53.61	51.42	-	-
8	49.69	46.46	50.45	51.17	49.96	51.91	49.72	48.29	-
9	50.11	46.88	49.96	49.25	50.07	47.68	49.87	51.29	50.42

表 8: 对焦分类在训练集上的初始正确率%

负-正	0	1	2	3	4	5	6	7	8
1	99.94	-	-	-	-	-	-	-	-
2	98.23	99.02	-	-	-	-	-	-	-
3	98.68	99.48	93.56	-	-	-	-	-	-
4	99.80	99.66	99.00	99.11	-	-	-	-	-
5	97.97	99.05	96.91	87.27	97.13	-	-	-	-
6	98.28	99.64	95.36	98.23	97.91	95.96	-	-	-
7	99.58	99.71	97.59	98.47	97.60	98.05	99.32	-	-
8	98.71	98.62	96.44	88.94	97.55	96.86	97.33	96.87	-
9	99.00	99.53	98.29	97.64	89.40	95.88	99.28	92.42	95.86

表 9: 对焦分类的锚点与最优锚点的相对距离%

负-正	0	1	2	3	4	5	6	7	8
1	-8.42	-	-	-	-	-	-	-	-
2	-3.89	12.47	-	-	-	-	-	-	-
3	-5.10	9.52	-1.38	-	-	-	-	-	-
4	-6.94	5.49	-6.85	-3.13	-	-	-	-	-
5	-8.83	7.10	-2.26	2.13	2.56	-	-	-	-
6	-5.05	8.30	-2.66	-1.09	2.16	0.53	-	-	-
7	-6.50	6.90	-3.95	-4.31	1.35	-3.18	-1.59	-	-
8	-4.55	12.35	-4.03	1.50	2.89	9.99	0.49	2.36	-
9	-5.15	5.00	-4.23	-5.05	-4.13	-1.24	-1.41	-0.07	-4.99

对焦回归的初始值选择有明确的理论指导，导致初始正确率较高，见表 8，平均为 97.31%。

由式 (16) 知，锚点 $\mathbf{c}$ 与最优锚点 $\hat{\mathbf{c}}$ 的之间绝对距离为 $\hat{b}/|\hat{\mathbf{w}}|$，相对距离可以用绝对距离与 $|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0|/2$ 之间的比值表示，即

$$\frac{2\hat{b}}{|\hat{\mathbf{w}}||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0|}. \tag{34}$$

训练集上的相对距离列在表 9，平均值为-0.3751%，可见锚点的选取十分精确。

8.3 收敛速度

用 2 个指标来衡量收敛速度：达到最高正确率的 99% 和最高正确率的 99.9% 所需要的迭代次数。之所以不用达到最高正确率所需的迭代次数，是因为在最高点附近几乎都会发生

表 10: 逻辑回归 99% 分位迭代次数

负-正	0	1	2	3	4	5	6	7	8
1	2	-	-	-	-	-	-	-	-
2	96	74	-	-	-	-	-	-	-
3	76	66	122	-	-	-	-	-	-
4	2	19	51	34	-	-	-	-	-
5	141	81	93	380	102	-	-	-	-
6	68	35	269	53	21	120	-	-	-
7	24	85	177	143	133	32	2	-	-
8	76	166	216	256	79	469	50	121	-
9	55	25	101	258	369	122	9	381	271

表 11: 逻辑回归 99.9% 分位迭代次数

负-正	0	1	2	3	4	5	6	7	8
1	268	-	-	-	-	-	-	-	-
2	1275	1222	-	-	-	-	-	-	-
3	1110	1200	1253	-	-	-	-	-	-
4	474	847	1202	970	-	-	-	-	-
5	1515	1379	1238	1481	1121	-	-	-	-
6	1120	1073	1502	1011	745	1487	-	-	-
7	691	1119	1406	1019	1021	1118	335	-	-
8	1113	1482	1514	1523	1336	1568	1289	1265	-
9	800	699	1133	1554	1792	1337	541	1733	1680

微小振荡，随机性强，难以正确反映算法的特性。

逻辑回归的 99% 分位迭代次数列在表 10，平均值为 122.78；逻辑回归的 99.9% 分位迭代次数列在表 11，平均值为 1168.02。

对焦分类的 99% 分位迭代次数列在表 12，平均值为 23.69；对焦分类的 99.9% 分位迭代次数列在表 13，平均值为 361.16.

由表 10~13易知，对焦分类的收敛速度比逻辑回归快很多。

逻辑回归的截距 b 的初值列在表 14，平均值为 0.0513；迭代 2000 次之后截距 b 的终值列在表 15，平均值为 0.0941；可以将截距 b 的初值与终值差的绝对值称为接近度，接近度越小越好，逻辑回归的接近度均值为 0.3710。

对焦回归的离心距 b 初值为 0，迭代 2000 次之后离心距 b 的值列在表 16，平均值为-0.0069，接近度平均值为 0.0737。与逻辑回归相比，对焦回归的初值更接近终值，更好。

表 12: 对焦分类 99% 分位迭代次数

负-正	0	1	2	3	4	5	6	7	8
1	0	-	-	-	-	-	-	-	-
2	2	0	-	-	-	-	-	-	-
3	2	0	41	-	-	-	-	-	-
4	0	0	0	0	-	-	-	-	-
5	8	0	8	263	17	-	-	-	-
6	3	0	108	0	6	16	-	-	-
7	0	0	12	0	13	13	0	-	-
8	0	1	16	123	18	25	1	29	-
9	0	0	3	6	61	27	0	198	46

表 13: 对焦分类 99% 分位迭代次数

负-正	0	1	2	3	4	5	6	7	8
1	0	-	-	-	-	-	-	-	-
2	102	41	-	-	-	-	-	-	-
3	95	108	325	-	-	-	-	-	-
4	19	30	100	69	-	-	-	-	-
5	124	154	51	1709	164	-	-	-	-
6	236	134	1303	150	189	414	-	-	-
7	112	55	340	184	1025	731	161	-	-
8	109	99	185	1243	418	296	97	628	-
9	285	68	90	691	418	1268	230	1228	774

表 14: 逻辑回归截距 b 初值

负-正	0	1	2	3	4	5	6	7	8
1	.8560	-	-	-	-	-	-	-	-
2	.2500	-.1306	-	-	-	-	-	-	-
3	-.2662	-.3966	-.3327	-	-	-	-	-	-
4	.4788	.2850	.0470	.8713	-	-	-	-	-
5	.3807	.6325	-.2973	-.3549	-.9653	-	-	-	-
6	.9041	-.1027	-.5404	.9929	-.2355	.3973	-	-	-
7	-.7932	-.5697	-.8572	.5080	.0098	-.3679	.2959	-	-
8	-.8116	.4658	.9810	-.0130	.1991	.9828	.7160	.8654	-
9	-.1333	-.4223	-.6506	-.9309	.1562	.8105	.1012	-.1354	-.5732

表 15: 逻辑回归 2000 次迭代后的截距 b

负-正	0	1	2	3	4	5	6	7	8
1	.4280	-	-	-	-	-	-	-	-
2	-.1410	.2158	-	-	-	-	-	-	-
3	-.4141	-.0962	.1157	-	-	-	-	-	-
4	.1230	.3888	.0514	.5935	-	-	-	-	-
5	-.8248	.4153	-.7692	-1.0776	-1.2000	-	-	-	-
6	.4611	.1806	-.0488	.7307	.0464	1.0269	-	-	-
7	-1.1317	-.3160	-.8410	-.1229	-.2266	-.2077	.0680	-	-
8	-.6578	1.0086	1.3943	.6378	.6047	2.5445	.8359	1.3909	-
9	-.3167	-.2194	-.5657	-.8159	.3086	1.1756	.0268	.5616	-1.1064

表 16: 对焦分类 2000 次迭代后的离心距 b

负-正	0	1	2	3	4	5	6	7	8
1	-.1480	-	-	-	-	-	-	-	-
2	-.0639	.2044	-	-	-	-	-	-	-
3	-.0833	.1574	-.0212	-	-	-	-	-	-
4	-.1123	.1007	-.1024	-.0523	-	-	-	-	-
5	-.1348	.1247	-.0352	.0285	.0403	-	-	-	-
6	-.0893	.1416	-.0423	-.0197	.0346	.0089	-	-	-
7	-.1194	.1226	-.0755	-.0772	.0189	-.0531	-.0294	-	-
8	-.0749	.1900	-.0563	.0212	.0447	.1246	.0086	.0421	-
9	-.0926	.0889	-.0756	-.0841	-.0518	-.0187	-.0226	-.0009	-.0772

表 17: 逻辑回归初始法向量模长

负-正	0	1	2	3	4	5	6	7	8
1	.0210	-	-	-	-	-	-	-	-
2	.0205	.0204	-	-	-	-	-	-	-
3	.0204	.0207	.0210	-	-	-	-	-	-
4	.0205	.0209	.0207	.0201	-	-	-	-	-
5	.0207	.0204	.0206	.0208	.0209	-	-	-	-
6	.0212	.0200	.0209	.0212	.0208	.0204	-	-	-
7	.0212	.0205	.0211	.0204	.0209	.0204	.0209	-	-
8	.0202	.0203	.0209	.0209	.0201	.0209	.0203	.0204	-
9	.0208	.0209	.0209	.0196	.0204	.0205	.0200	.0202	.0214

表 18: 逻辑回归 2000 次迭代后的法向量模长

负-正	0	1	2	3	4	5	6	7	8
1	2.7713	-	-	-	-	-	-	-	-
2	3.5602	3.8066	-	-	-	-	-	-	-
3	3.5543	3.5001	3.6743	-	-	-	-	-	-
4	3.1414	3.2576	3.5991	3.6558	-	-	-	-	-
5	4.0163	3.8027	3.7661	4.5227	4.1154	-	-	-	-
6	3.6352	3.4458	4.2158	3.7039	3.7896	3.8219	-	-	-
7	3.2619	3.6135	3.5932	3.5840	4.1163	4.0012	3.3663	-	-
8	3.3878	3.8553	3.9169	4.1160	4.0820	4.1455	3.7295	3.8305	-
9	3.2914	3.4038	3.7502	3.8728	5.0291	4.2215	3.5780	4.2077	4.2341

8.4 法向量模长

逻辑回归初始法向量模长列在表 17，平均值为 0.0206；逻辑回归在 2000 次迭代后的法向量模长列在表 18，平均值为 3.7677。对焦分类初始法向量模长列在表 19，平均值为 0.6546；对焦分类在 2000 次迭代后的法向量模长列在表 20，平均值为 0.6964，相对于初始值，仅增长了 6.39%

观察这些模长数据可知，逻辑回归的法向量模长较大；对焦分类的法向量模长较小，增长缓慢，极有可能是有界的，支持了定理 3。

表 19: 对焦分类初始法向量模长

负-正	0	1	2	3	4	5	6	7	8
1	.3967	-	-	-	-	-	-	-	-
2	.5333	.5923	-	-	-	-	-	-	-
3	.5368	.5845	.6829	-	-	-	-	-	-
4	.4805	.5300	.6338	.5710	-	-	-	-	-
5	.6617	.6209	.6628	.9628	.7283	-	-	-	-
6	.5277	.5280	.7594	.5586	.6998	.6869	-	-	-
7	.4724	.5587	.5589	.5735	.7453	.6713	.5296	-	-
8	.5345	.6299	.7802	.8272	.6909	.9666	.6410	.6261	-
9	.4856	.5647	.6077	.6274	1.2074	.7963	.6358	1.0045	.7839

表 20: 对焦分类 2000 次迭代后的法向量模长

负-正	0	1	2	3	4	5	6	7	8
1	.4648	-	-	-	-	-	-	-	-
2	.5842	.6473	-	-	-	-	-	-	-
3	.5841	.6444	.7002	-	-	-	-	-	-
4	.5188	.6487	.6314	.6370	-	-	-	-	-
5	.6730	.7267	.6897	.8589	.7653	-	-	-	-
6	.6214	.6004	.8046	.6716	.7462	.7694	-	-	-
7	.5788	.6618	.7128	.6858	.6974	.7457	.6515	-	-
8	.5867	.6459	.7270	.7805	.7135	.8035	.7423	.7464	-
9	.5815	.6699	.7236	.6969	1.0099	.8020	.6823	.8944	.8078

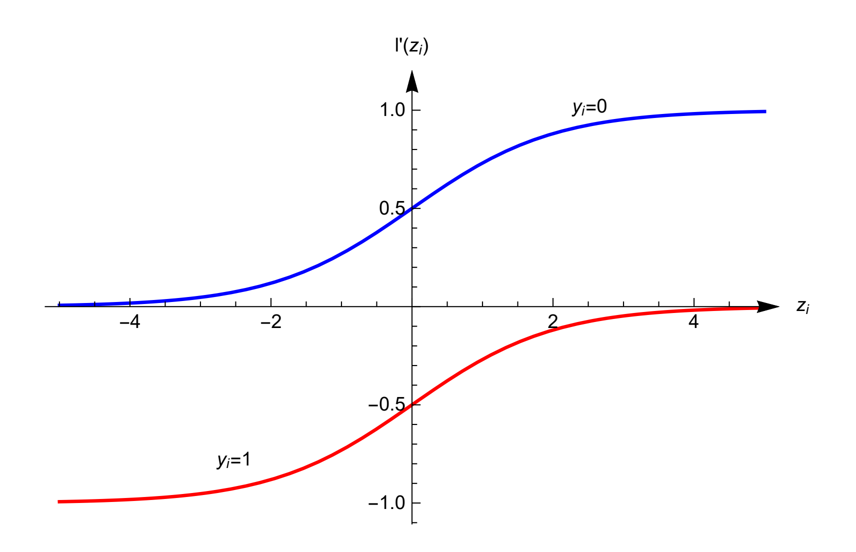


图 7: 逻辑回归：单样本损失函数的导数 $l(z_i)$。

9 总结

本文设计的对焦分类方法解决了逻辑回归面临的 2 个困难：法向量模长过大、初值不准，并提供了理论证明和直观的几何解释性。虽然无法保证控制法向量模长一定能够缓解过拟合现象，但仍然可以在有过拟合现象的数据集测试对焦分类的表现。

后续工作可以去证明定理 3，将对焦分类并行化。

10 附录

为证明逻辑回归的法向无限定理和法向有界定理，先考查损失函数的导数。

由式 (3) 得

$$l'(z_i) = \begin{cases} \sigma(z_i), & \text{如果 } y_i = 0, \\ \sigma(z_i) - 1, & \text{如果 } y_i = 1. \end{cases} \tag{35}$$

$l'(z_i)$ 的图像如图 2所示。对正样本 $\mathbf{x}_i$，当 $z_i > 0$ 时 $-0.5 < l'(z_i) < 0$，当 $z_i < 0$ 时 $-1 < l'(z_i) < -0.5$，当 $z_i = 0$ 时 $l'(z_i) = -0.5$；对负样本 $\mathbf{x}_i$，当 $z_i > 0$ 时 $0.5 < l'(z_i) < 1$，当 $z_i < 0$ 时 $0 < l'(z_i) < 0.5$，当 $z_i = 0$ 时 $l'(z_i) = 0.5$。

定理 4 (法向无限). *样本集 D 线性可分时，逻辑回归最优分隔平面的法向量 $\hat{\mathbf{w}}$ 的模长是 $+\infty$。*

证: 根据线性可分的定义，存在向量 $\mathbf{w}_1$ 和标量 c_1，$\forall \mathbf{x}_i \in D$ 满足

$$z_{i1} = \begin{cases} \mathbf{w}_1^T(\mathbf{x}_i - \mathbf{c}_1) < 0, & \text{如果 } y_i = 0, \\ \mathbf{w}_1^T(\mathbf{x}_i - \mathbf{c}_1) \geq 0, & \text{如果 } y_i = 1, \end{cases}$$

令 $\mathbf{w}_2 = 2\mathbf{w}_1$，那么有

$$z_{i2} = \mathbf{w}_2^T(\mathbf{x}_i - \mathbf{c}_1) = 2\mathbf{w}_1^T(\mathbf{x}_i - \mathbf{c}_1) = z_{i1}, i = 1, 2, \ldots, m,$$

因此, 对任意的正整数 $1 \leq i \leq m$，有

$$l(z_{i2}) = \begin{cases} -\ln(1 - \sigma(2z_{i1})) < -\ln(1 - \sigma(z_{i1})), & \text{如果 } z_{i1} < 0, \\ -\ln(\sigma(2z_{i1})) \leq -\ln(\sigma(z_{i1})), & \text{如果 } z_{i1} \geq 0. \end{cases}$$

即 $L(\mathbf{w}_2, b) < L(\mathbf{w}_1, b)$。按照这个每次模长加倍的方法推下去，就得到 $|\hat{\mathbf{w}}| = +\infty$。

[证毕]

定理 5 (法向有界). *样本集 D 线性不可分时，逻辑回归最优分隔平面的法向量有界。*

证: 假设满足式 (4) 的最优分隔平面的点法式方程为

$$\hat{\mathbf{w}}^T(\mathbf{x} - \hat{\mathbf{c}}_1) = 0.$$

令 $\mathbf{n} = \hat{\mathbf{w}}/|\hat{\mathbf{w}}|$，显然 $|\mathbf{n}| = 1$。假设样本集 D 线性不可分，从而存在指标 i 使得

$$z_i = \mathbf{n}^T(\mathbf{x}_i - \mathbf{c}_1) \geq 0, \text{且 } y_i = 0, \tag{36}$$

或者

$$z_i = \mathbf{n}^T(\mathbf{x}_i - \mathbf{c}_1) < 0, \text{且 } y_i = 1. \tag{37}$$

将指标集合记为

$$\begin{aligned} I_0 &= \{i | z_i < 0 \text{ 且 } y_i = 0, \quad 1 \leq i \leq m\}, \\ I_1 &= \{i | z_i > 0 \text{ 且 } y_i = 1, \quad 1 \leq i \leq m\}, \\ J_0 &= \{j | z_j > 0 \text{ 且 } y_j = 0, \quad 1 \leq j \leq m\}, \\ J_1 &= \{j | z_j < 0 \text{ 且 } y_j = 1, \quad 1 \leq j \leq m\}, \\ K_0 &= \{k | z_k = 0, \quad 1 \leq k \leq m\}, \end{aligned} \tag{38}$$

指标集合上的损失函数分别记为

$$\begin{aligned} L_{I_0}(\mathbf{n}) &= \frac{1}{m}\sum_{i \in I_0} l(z_i), \quad L_{I_1}(\mathbf{n}) = \frac{1}{m}\sum_{i \in I_1} l(z_i), \\ L_{J_0}(\mathbf{n}) &= \frac{1}{m}\sum_{j \in J_0} l(z_j), \quad L_{J_1}(\mathbf{n}) = \frac{1}{m}\sum_{j \in J_1} l(z_j), \\ L_{K_0}(\mathbf{n}) &= \frac{1}{m}\sum_{k \in K_0} l(z_k) \end{aligned} \tag{39}$$

由式 (3) 知，

$$L(\mathbf{n},b)=L_{I_0}(\mathbf{n})+L_{I_1}(\mathbf{n})+L_{J_0}(\mathbf{n})+L_{J_1}(\mathbf{n})+L_{K_0}(\mathbf{n}).$$

由式 (36)(37) 知 $I_0\cup I_1\cup K_0$ 和 $J_0\cup J_1\neq\phi$ 是非空集合，为论证方便，这里仅考虑 I_0、I_1、J_0、J_1、K_0 均为非空集合的一般情形，其它特殊情形可做类似证明。

令 $\lambda>1$ 为正实数，δ 为正无穷小量。接下来寻找 λ 的取值范围，使得

$$\begin{aligned}
&L((\lambda+\delta)\mathbf{n},b)-L(\lambda\mathbf{n},b)\\
&=\ L_{I_0}((\lambda+\delta)\mathbf{n})-L_{I_0}(\lambda\mathbf{n})+L_{I_1}((\lambda+\delta)\mathbf{n})-L_{I_1}(\lambda\mathbf{n})+L_{J_0}((\lambda+\delta)\mathbf{n})\\
&\quad -L_{J_0}(\lambda\mathbf{n})+L_{J_1}((\lambda+\delta)\mathbf{n})-L_{J_1}(\lambda\mathbf{n})+L_{K_0}((\lambda+\delta)\mathbf{n})-L_{K_0}(\lambda\mathbf{n})\\
&>\ 0
\end{aligned}\tag{40}$$

成立。

由式 (38)(3) 知，对 $\forall k\in K_0$，有 $z_k=0$，$l((\lambda+\delta)z_k)=l(\lambda z_k)=\ln(2)$。由式 (39) 知，

$$L_{K_0}((\lambda+\delta)\mathbf{n})-L_{K_0}(\lambda\mathbf{n})=0.\tag{41}$$

对 $\forall j\in J_0$，由式 (35) 知，$l'(z_j)$ 的值从 $1/2$ 严格单调递增至 1，从而有

$$l((\lambda+\delta)z_j)-l(\lambda z_j)>l'(\lambda z_j)\delta z_j>\frac{1}{2}\delta z_j,$$

$$L_{J_0}((\lambda+\delta)\mathbf{n})-L_{J_0}(\lambda\mathbf{n})>\frac{1}{2}\delta\sum_{j\in J_0}z_j.\tag{42}$$

对 $\forall j\in J_1$，由式 (35) 知，$l'(z_j)$ 的值从 -1 严格单调递增至 $-1/2$，从而有

$$l((\lambda+\delta)z_j)-l(\lambda z_j)>l'(\lambda z_j)\delta z_j>-\frac{1}{2}\delta z_j,$$

$$L_{J_1}((\lambda+\delta)\mathbf{n})-L_{J_1}(\lambda\mathbf{n})>-\frac{1}{2}\delta\sum_{j\in J_1}z_j.\tag{43}$$

令

$$E_0=\frac{1}{2}(\sum_{j\in J_0}z_j-\sum_{j\in J_1}z_j)/(-\sum_{i\in I_0}z_i+\sum_{i\in I_1}z_i),$$

由 I_0、I_1、J_0、J_1 的定义知 $E_0>0$。这里需要假设 $E_0<1$，它可以模糊地理解为"样本集中被分错的样本数量小于被分对的样本数量的 2 倍"，显然是一个合理的假设。由式 (35) 知，在 $y_i=0$ 时，$l'(z_i)$ 在定义域 $(-\infty,0)$ 上的值从 0 严格单调递增至 $1/2$，因此对 $\forall\lambda>\lambda_0=-\sigma^{-1}(E_0)>0$ 和 $\forall i\in I_0$ 均有 $l'(\lambda z_i)<E_0$。

对 $\forall i\in I_0$，由式 (3)(35) 知

$$0>l((\lambda+\delta)z_i)-l(\lambda z_i)>l'(\lambda z_i)\delta z_i,$$

进而，对 $\forall\lambda > \lambda_0$ 有

$$L_{I_0}((\lambda+\delta)\mathbf{n}) - L_{I_0}(\lambda\mathbf{n}) > \delta\sum_{i\in I_0} l'(\lambda z_i)z_i > \delta E_0\sum_{i\in I_0} z_i. \tag{44}$$

由式 (35) 知，在 $y_i = 1$ 时，$l'(z_i)$ 在定义域 $(0,+\infty)$ 上的值从 $-1/2$ 严格单调递增至 0，因此对 $\forall\lambda > \lambda_1 = \sigma^{-1}(1-E_0) > 0$ 和 $\forall i \in I_1$ 有

$$-E_0 < l'(\lambda z_i) < 0.$$

对 $\forall i \in I_1$，由式 (3)(35) 知

$$0 > l((\lambda+\delta)z_i) - l(\lambda z_i) > l'(\lambda z_i)\delta z_j,$$

进而，对 $\forall\lambda > \lambda_1$ 有

$$L_{I_1}((\lambda+\delta)\mathbf{n}) - L_{I_1}(\lambda\mathbf{n}) > \delta\sum_{i\in I_1} l'(\lambda z_i)z_i > -\delta E_0\sum_{i\in I_1} z_i. \tag{45}$$

综合式 (40)-(45) 得知，当 $\lambda > \max\{\lambda_0,\lambda_1\}$ 时，有

$$\begin{aligned}
& L((\lambda+\delta)\mathbf{n},b) - L(\lambda\mathbf{n},b) \\
> \ & \delta E_0\sum_{i\in I_0} z_i - \delta E_0\sum_{i\in I_1} z_i + \frac{\delta}{2}\sum_{j\in J_0} z_j - \frac{\delta}{2}\sum_{j\in J_1} z_j + 0 \\
> \ & \delta E_0(\sum_{i\in I_0} z_i - \sum_{i\in I_1} z_i) + \frac{\delta}{2}(\sum_{j\in J_0} z_j - \sum_{j\in J_1} z_j) \\
\geq \ & \frac{\delta}{2}\frac{\sum_{j\in J_0} z_j - \sum_{j\in J_1} z_j}{2-\sum_{i\in I_0} z_i + \sum_{i\in I_1} z_i}(\sum_{i\in I_0} z_i - \sum_{i\in I_1} z_i) + \frac{\delta}{2}(\sum_{j\in J_0} z_j - \sum_{j\in J_1} z_j) \\
= \ & 0.
\end{aligned} \tag{46}$$

式 (46) 意味着 $L(\lambda\mathbf{n},b)$ 在 $\lambda \in (\max\{\lambda_0,\lambda_1\},+\infty)$ 严格单调递增，从而最分优分隔平面的法向量模长 $|\hat{\mathbf{w}}| = \hat{\lambda}|\mathbf{n}| = \hat{\lambda} \leq \max\{\lambda_0,\lambda_1\}$，即 $\hat{\mathbf{w}}$ 有界。

[证毕]

参考文献

[1] 周志华，机器学习，p57-60，清华大学出版社，2016.4

[2] https://www.csie.ntu.edu.tw/ cjlin/liblinear/

[3] T. Hastie, R. Tibshirani, and J. Friedman. The Elements of Statisti- cal Learning, Second Edition: Data Mining, Inference, and Prediction. Springer Series in Statistics. Springer, 0002-2009. corr. 3rd edition, Feb. 2009.

[4] G. Andrew and J. Gao. Scalable training of l1-regularized log-linear models. In Proceedings of the 24th international conference on Machine learning, ICML '07, pages 33–40, New York, NY, USA, 2007. ACM.

[5] C.-J. Lin and J. J. Mor e. Newton's method for large bound-constrained optimization problems. SIAM J. on Optimization, 9(4):1100–1127, Apr. 1999.

[6] C.-J. Lin, R. C. Weng, and S. S. Keerthi. Trust region newton method for logistic regression. Journal of Machine Learning Research, 9:627–650, 2008.

[7] G.-X. Yuan, K.-W. Chang, C.-J. Hsieh, and C.-J. Lin. A comparison of optimization methods and software for large-scale l1-regularized linear clas- sification. J. Mach. Learn. Res., 11:3183–3234, Dec. 2010

[8] S. Perkins and J. Theiler. Online feature selection using grafting. In In In- ternational Conference on Machine Learning, pages 592–599. ACM Press, 2003.

[9] M. J. Streeter and H. B. McMahan. Less regret via online conditioning. CoRR, abs/1002.4862, 2010.

[10] G.-X. Yuan, C.-H. Ho, and C.-J. Lin. An improved glmnet for l1-regularized logistic regression. Journal of Machine Learning Research, 13:1999–2030, 2012.

[11] J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for gener- alized linear models via coordinate descent. Journal of Statistical Software, 33(1):1–22, 2010.

[12] http://yann.lecun.com/exdb/mnist/